The IBM Speech Activity Detection System for the DARPA RATS Program

George Saon, Samuel Thomas, Hagen Soltau, Sriram Ganapathy and Brian Kingsbury

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
gsaon@us.ibm.com

Abstract

We present the IBM speech activity detection system that was fielded in the phase 2 evaluation of the DARPA RATS (robust automatic transcription of speech) program. Key ingredients of the system are: multi-pass HMM Viterbi segmentation, fusion of multiple feature streams, file-based and speech-based normalization schemes, the use of regular and convolutional deep neural networks, and model fusion through frame-level score combination of channel-dependent models. These techniques were instrumental in achieving a 1.4% equal error rate on the RATS phase 2 evaluation data. → absolute ?

reduction ?

Index Terms: speech activity detection, robust speech recognition

1. Introduction

The goal of the DARPA RATS program is to develop techniques for performing speech activity detection (SAD), language identification (LID), speaker identification (SID) and keyword search (KWS) in multiple languages on degraded audio signals transmitted over communication channels that are extremely noisy and/or highly distorted [1]. The speech activity detection task deals with determining whether a signal contains speech or is just comprised of background noise or music. The segmented speech regions can be send downstream to the other components (LID, SID and KWS) for further processing as done in [2],[3] or can be directly used by analysts. Given its importance in the context of this program, SAD is evaluated in isolation of the other components. The performance metric used in this paper is the equal error rate which is defined as the point where the probability of miss (P_{Miss}) coincides with the probability of false accept (P_{FA}). These two quantities are defined as the duration of missed speech over the duration of total speech and the duration of false accept (or inserted) speech over the duration of total non-speech, respectively.

The paper is organized as follows: in section 2 we describe the system architecture, feature extraction, normalization and segmentation models; in section 3 we present some experimental results, and in section 4 we summarize our findings and propose future directions.

2. System overview

The operation of our system may be broken down into three stages depicted in Figure 1: (1) channel detection with 8 channel-dependent Gaussian mixture models trained with maximum likelihood on a fusion of PLP and voicing features (see 2.3.1), (2) speech/non-speech HMM Viterbi segmentation using channel-dependent deep neural networks (DNNs) trained on a fusion of PLP, voicing and rate-scale features with file-based mean and variance normalization (see 2.3.3) and (3) speech/non-speech HMM Viterbi segmentation using a frame-level score combination of three sets of channel-dependent neural networks: (i) the models from (2), (ii) DNNs trained on a fusion of PLP, voicing and FDLP features with speech-based mean and variance normalization (see 2.3.2) and (iii) deep convolutional neural nets (CNNs) trained on log-mel spectra with speech-based mean and variance normalization (see 2.3.4). The speech segments needed for speech-based normalization are hypothesised in pass (2).

expand

expansion

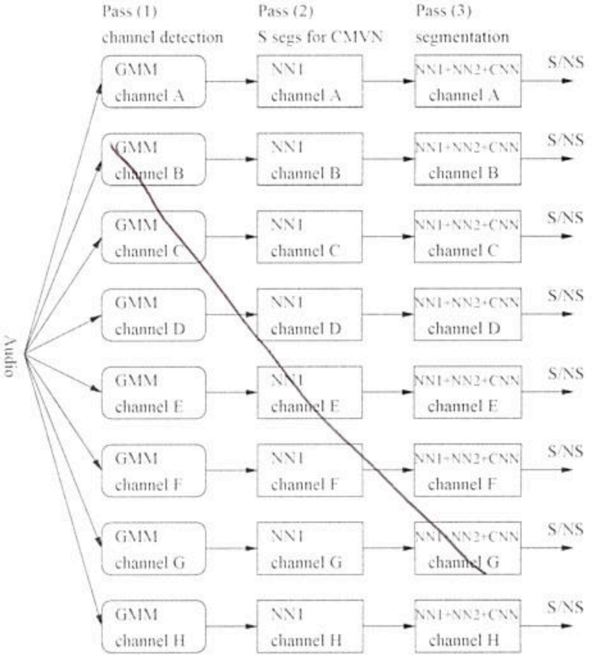

Figure 1: System diagram.

2.1. HMM Viterbi segmentation

We propose to treat the segmentation problem as a simple ASR decoding problem with a three word vocabulary (S, NS, NT) similar to [4]. The HMM topology used for Viterbi decoding is shown in Figure 2. All 5 states for a given "word" share the same output distribution. Analogous to the LM score, the segment insertion penalty controls the number (and duration) of the segments. The tradeoff between missed speech and inserted speech is controled by adding a fixed threshold to the S scores for every frame. The frame-level scores are scaled by an acoustic weight of 0.03 for all the experiments. Following the decoding, the boundaries of the hypothesized speech segments are extended by an additional 0.1 seconds to capture low energy speech as suggested in [5].

Model	Data	DEV1	DEV2
PLP+v+FDLP-F	1/10th	1.52	2.25
PLP+v+FDLP-F	all	1.16	2.06
PLP+v+FDLP-S	1/10th	1.36	2.24
PLP+v+FDLP-S	all	1.01	1.96
PLP+v+RS-F	1/10th	1.51	2.26
PLP+v+RS-F	all	1.14	2.01

Table 3: Equal error rates (%) on DEV1 and DEV2 for neural networks trained on subsampled and entire data on various feature streams (-F,-S stand for file-based and speech-based normalization, respectively).

is desirable for system combination. Model fusion was also preferred over additional feature fusion because adding more feature streams results in prohibitive disk space and training time requirements when training on all the available data. In Figure 3 we show the ROC curves for the individual networks and the model fusion on DEV1 and DEV2.

increase line width? and shape of pts?

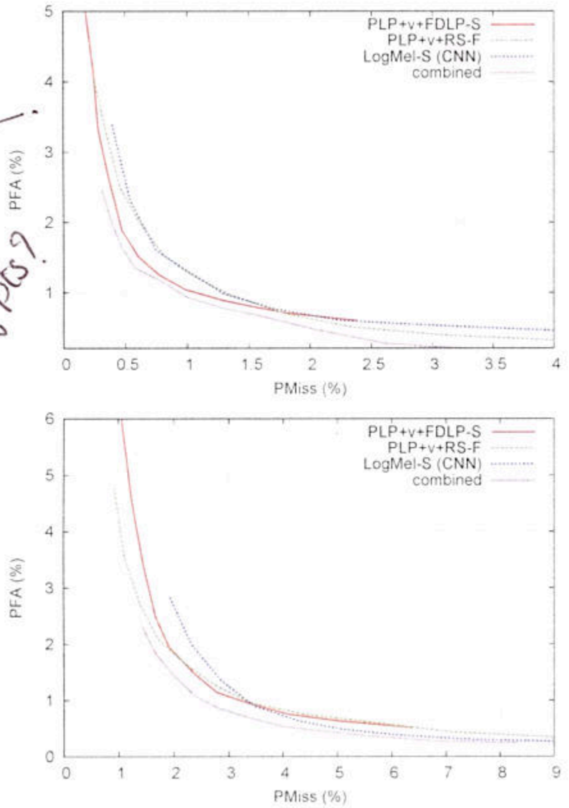

Figure 3: ROC curves for individual networks and model fusion on DEV1 (top) and DEV2 (bottom).

4. Conclusion

We have presented the speech activity detection system developed by IBM for the RATS phase 2 evaluation. This system achieved equal error rates of 1.0% on DEV1, 1.7% on DEV2 and 1.4% on the phase 2 evaluation data. The techniques that were instrumental in reaching this level of performance are: the use of regular and convolutional channel-dependent neural networks, combining multiple feature streams that differ in type and normalization, training on all of the available data, and model fusion by combining the frame-level scores of neural networks that differ in type and input features. Future work will address supervised and unsupervised adaptation to unseen channels.

5. Acknowledgments

The authors wish to thank Tomas Beran for help with software optimization. This work was supported in part by Contract No. D11PC20192 DOI/NBC under the RATS program. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

6. References

[1] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. Odyssey*, 2012.

[2] K. Han, S. Ganapathy, M. Li, M. Omar, and S. Narayanan, "TRAP language identification system for RATS phase II evaluation," in *Proc. Interspeech*, 2013, Submitted.

[3] L. Mangu, H. Soltau, H.-K. Kuo, B. Kingsbury, and G. Saon, "Exploiting diversity for spoken term detection," in *Proc. of ICASSP*, 2013.

[4] G. Saon, G. Zweig, B. Kingsbury, L. Mangu, and U. Chaudhari, "An architecture for rapid decoding of large vocabulary conversational speech," in *Proc. Eurospeech*, 2003.

[5] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. Interspeech*, 2012.

[6] C.-P. Chen and J.A. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 1, 2007.

[7] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, 2002.

[8] M. Athineos and D. Ellis, "Autoregressive modelling of temporal envelopes," *IEEE Transactions on Signal Processing*, vol. 55, no. 11, 2007.

[9] S. Ganapathy, *Signal Analysis using Autoregressive models of amplitude modulaton*, Ph.D. thesis, Johns Hopkins University, 2012.

[10] T. Chi, P. Ru, and S.A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, 2005.

[11] S. Nemala, K. Patil, and M. Elhilali, "A multistream feature framework based on bandpass modulation filtering for robust speech recognition," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 2, 2013.

[12] F. Seide, G. Li, X. Chen, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, 2011.

[13] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Handwritten digit recognition with a back-propagation network," in *Proc. NIPS*, 1990.

[14] H. Soltau, "Acoustic modeling for the DARPA RATS program," in *Proc. Interspeech*, 2013, Submitted.

[15] S. Sadjadi and J. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letter*, vol. 20, no. 3, 2013.